Sum of Minimum Frequencies as a Measure of Corpus Similarity

Alexander Piperski (Russian State University for the Humanities, Moscow, Russia)

This paper introduces a new measure of corpus similarity called SMF (Sum of Minimum Frequencies). This measure is conceived as an expected proportion of shared words in two large samples with replacement drawn from the corpora under comparison. SMF is shown to achieve comparable results to other word frequency measures of corpus similarity, namely, Spearman and χ^2. The paper also discusses some possible modifications of the SMF metric involving trimming vocabulary lists and applying a power function to corpus frequencies.

1. Introduction

The problem of what is corpus similarity and how to measure it has been widely discussed within corpus linguistics, starting with Kilgarriff (1997) and Kilgarriff & Rose (1998). Since then, many measures of corpus similarity have been proposed (see Kilgarriff 2001, Kilgarriff 2009, Fothergill, Cook, & Baldwin 2016 among others), based on word frequencies, perplexity, and topic modelling. However, no definitive measure for corpus similarity has yet been found. In this paper, I propose and evaluate a new corpus similarity measure based on word frequencies. This measure is easy both to compute and to interpret.

2. Sum of Minimum Frequencies (SMF)

Let us assume that we have two corpora C_1 and C_2 we want to compare to each other. These corpora use a vocabulary $W = \{w_1, w_2, \ldots, w_n\}$. The relative frequencies of a word w in these two corpora are $f_1(w)$ and $f_2(w)$. Now let us take two samples with replacement S_1 and S_2 from C_1 and C_2, respectively. If S_1 and S_2 contain a large proportion of shared words, C_1 and C_2 can be considered similar, and if this proportion is small, the corpora are dissimilar.

For instance, let us take three corpora C_1, C_2, and C_3 with a union of their vocabularies comprising five words $\{a, b, c, d, e\}$. The frequencies of these words in the three corpora are given in Table 1:

	C_1	C_2	C_3
$w_1 = a$	0.4	0.5	0.2
$w_2 = b$	0.3	0.2	0.1
$w_3 = c$	0.2	0.2	0.1
$w_4 = d$	0.1	0.0	0.1
$w_5 = e$	0.0	0.1	0.5

Table 1. Three corpora for comparison: a toy example

Samples of 100 words drawn with replacement from each corpus would look like the following:

S_1: 37 *a*'s, 29 *b*'s, 25 *c*'s, and 9 *d*'s
S_2: 56 *a*'s, 16 *b*'s, 17 *c*'s, and 11 *e*'s
S_3: 20 *a*'s, 6 *b*'s, 7 *c*'s, 11 *d*'s, and 56 *e*'s

Samples S_1 and S_2 have 70% of words in common (37 *a*'s, 16 *b*'s, and 17 *c*'s out of 100 words), S_1 and S_3 have 44% of words in common (20 *a*'s, 6 *b*'s, 7 *c*'s, and 11 *d*'s), as well as S_2 and S_3 (20 *a*'s, 6 *b*'s, 7 *c*'s, and 11 *e*'s). This shows that C_1 and C_2 are more similar to each other than any of them is to C_3, which is the conclusion we would expect looking at the frequency lists.

In order to calculate the similarity score, we do not actually need to draw random samples, since frequency lists provide sufficient information to estimate the expected value of the sample intersection size. It is simply the **Sum of Minimum Frequencies (SMF)** for all the words in the vocabulary:

$$SMF = \sum_{i=1}^{n} min\big(f_1(w_i), f_2(w_i)\big)$$

Since $f_x(w)$ are relative frequencies by definition, SMF can range between 0 and 1, where 0 stands for absolute dissimilarity (it means that the two corpora do not share a single vocabulary item), and 1 stands for complete identity of the frequency lists for the two corpora.

To my knowledge, such a method was first used in biodiversity studies (Renkonen 1938). Drawing random samples and comparing them was proposed by Labbé & Labbé (2012). Shaikevich (2015) also compares corpora along similar lines, but he takes only *k* most frequent words from the corpora for comparison.

3. Evaluation

The standard way to evaluate a similarity metric is to use Kilgarriff's (2001) Known-Similarity Corpora (KSC) approach. In order to evaluate the SFM metric, I ran a series of experiments on the same set of BNC subcorpora as Kilgarriff had done:

> Accountancy (acc); The Art Newspaper (art); British Medical Journal (bmj); Environment Digest (env); The Guardian (gua); The Scotsman (sco); Today (tod).

For each pair of these seven corpora, a KSC-set comprising 11 corpora was generated (e.g., 0% acc and 100% gua, 10% acc and 90% gua, … , and 100% acc and 0% gua, no text fragment appearing in more than one corpus), and for these 21 KSC-sets all 660 KSC similarity judgments were checked against the a priori gold standard judgments. The percentage of times the SMF measure agreed with the gold standard is shown in Table 2, the mean percentage of agreement being 93.57 and the median being 96.21:

	acc	art	bmj	env	gua	sco	tod
acc		94.39	90.76	98.33	**85.00**	75.76	92.27
art			96.52	99.70	**95.45**	89.85	98.03

bmj	99.24	**96.06**	97.88	98.03
env		**99.24**	98.48	99.39
gua			80.30	95.30
sco				84.85

Table 2. The percentages of KSC gold standard judgments correctly captured by SMF

These results can be compared with other corpus similarity measures. Kilgarriff (2001) provides the results for five measures tested on four KSC-sets, see Table 3 below. The results of SMF on these four sets are highlighted in bold in Table 2.

	spear	χ^2	closed	type 1	type 2
KSC-set					
acc_gua	93.33	91.33	82.22	81.11	80.44
art_gua	95.60	93.03	84.00	83.77	84.00
bmj_gua	95.57	97.27	88.77	89.11	88.77
env_gua	99.65	99.31	87.07	84.35	86.73

Table 3. The percentages of KSC gold standard judgments correctly captured by five corpus similarity measures, reproduced from Kilgarriff (2001: 127)

SMF outperforms the three perplexity measures ("closed", "type 1", and "type 2") in all four test cases. It comes extremely close to Spearman ("spear") and χ^2 in three out of four test cases, outperforming Spearman on the bmj_gua set and outperforming χ^2 on the art_gua set. This shows that the SMF metric achieves results comparable to the corpus similarity measures that were deemed best by Kilgarriff (2001).

4. Modifications of the SMF metric

A possible drawback of SMF is that it gives either too much or too little weight to high-frequency items, such as articles, pronouns, modal verbs, etc. In order to account for this, I apply different types of transformations to test whether increasing the weight of high- or low-frequency items would provide a performance gain.

One option to think of is to trim the frequency distributions from above, i.e. to exclude the union of the words that occupy top *k* positions in each corpus compared, thus removing some number of words between *k* to 2*k* from the vocabulary. However, it turns out that trimming from above does not bring any systematic improvement. The median percentage of agreement with the gold standard for the same 21 KSC-sets ranges between 95.30 and 96.36 (see Figure 1 below), but this is definitely not enough to say that trimming some positive number *k* of words from above is preferable to non-trimming, which yields a median of 96.21%:

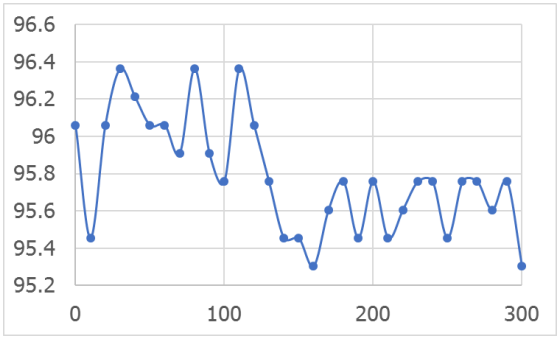

Figure 1. SMF with trimming from above (i.e. removing the union of *k* most frequent words from both corpora): median scores on 21 KSC-sets

Another possible option is to apply a power function to corpus frequencies. From the original frequency distribution $\{f(w_1), f(w_2), \ldots, f(w_n)\}$ we get the transformed distribution $\{f'(w_1), f'(w_2), \ldots, f'(w_n)\}$ using the following formula:

$$f'^{(w_i)} = \frac{f^p(w_i)}{\sum_{j=1}^{n} f^p(w_j)}$$

Setting p below 1 increases the weight of low-frequency items (in the extreme case where $p = 0$, all words in a corpus have equal impact on the outcome regardless of their frequency), whereas setting p above 1 increases the weight of high-frequency items. Testing different values of p between 0 and 2 with a step of 0.1 shows that the median for our 21 KSC-sets is the same at p = 1.0, 1.1, 1.2, and 1.3 and equals 96.21% (see Figure 2).

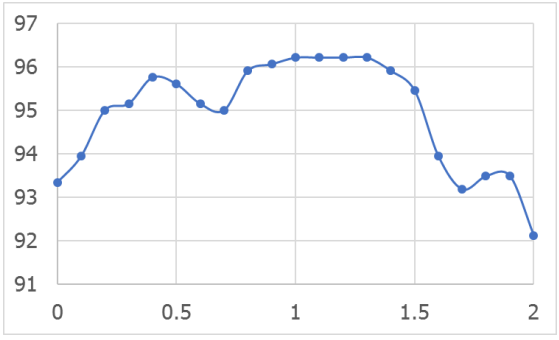

Figure 2. SMF with a power function applied: median scores on 21 KSC-sets

Interestingly, setting p to 1.1 or to 1.2 improves the performance on the four KSC-sets that are directly comparable to Kilgarriff's (2001) results, so that SMF even outperforms both Spearman and χ^2 two times out of four.

Another possible way of improving the performance of the SMF metric might be trimming the data from below, i.e. leaving only the union of k most frequent words from both corpora for comparison, thus neglecting the infrequent items. This method was first proposed by Shaikevich (2015). Let V be the union of k most frequent words from both corpora. The set $V = \{v_1, v_2, \ldots, v_m\}$ contains some number of words m between k and $2k$, and the formula for this version of SMF is as follows:

$$SMF(k \text{ most frequent}) = \frac{\sum \min\left(f_1(v_i), f_2(v_i)\right)}{\sum \max\left(f_1(v_i), f_2(v_i)\right)}$$

This tweak makes the SMF even more accurate. Figure 3 shows that the best results are achieved with *k* around 1000:

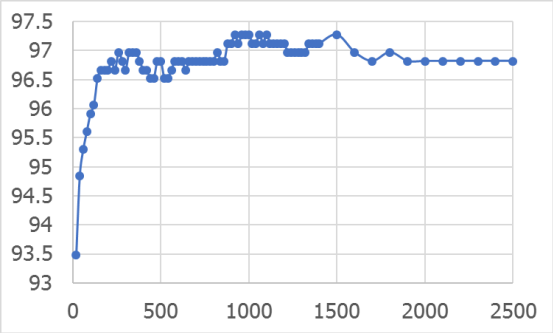

Figure 3. SMF with trimming from below
(i.e. leaving only the union of *k* most frequent words from both corpora)

However, the performance gain achieved by the modified versions of the SMF metric might be due to the nature of the test data. The effect of various tweaks on the accuracy of the SMF has to be tested more thoroughly. For this reason, it is sensible to stick to the simplest version of SMF for the time being, since it also has the advantage of being easily interpretable in a way described in Section 2.

5. Conclusion

The Sum of Minimum Frequencies (SMF) metric presented is this paper is yet another corpus similarity measure based on word frequencies. It outperforms perplexity measures and achieves results comparable to those of Spearman's rank correlation coefficient and χ^2. Some modifications applied to this metric may eventually lead to SMF outperforming the other two measures. However, even the simplest version of the SMF, outlined in Section 2, performs quite well and has an advantage of being easily interpretable as a proportion of shared words in two large samples with replacement drawn from the corpora under comparison.

Further directions of research include a more in-depth study of the modified versions of SMF, confronting SMF with the keyword-based corpus similarity measure outlined in Kilgarriff (2009) and implemented in SketchEngine, and comparing SMF to other corpus similarity measures using languages other than English.

References

Fothergill, R., Cook, P., & Baldwin, T. (2016). Evaluating a topic modelling approach to measuring corpus similarity. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia (pp. 273–279).

Kilgarriff, A. (1997). Using word frequency lists to measure corpus homogeneity and similarity between corpora. ACL W97-0122.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), 97–133. doi: 10.1075/ijcl.6.1.05kil

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, Liverpool, UK.

Kilgarriff, A., & Rose, T. (1998). Measures for corpus similarity and homogeneity. ACL W98-1506.

Labbé, C., & Labbé, D. (2012). Duplicate and fake publications in the scientific literature: how many SCIgen papers in computer science?. *Scientometrics*, 94(1), 379–396. doi: 10.1007/s11192-012-0781-y

Renkonen, O. (1938). Statistisch-ökologische Untersuchungen über die terrestrische Käferwelt der finnischen Bruchmoore. *Annales zoologici Societatis zoologicae-botanicae Fennicae Vanamo*, 6, 1–231.

Shaikevich, A. Ya. (2015). Mery leksičeskogo sxodstva častotnyh slovarej [Measures of lexical similarity for frequency dictionaries]. In *Proceedings of the International Conference "Corpus Linguistics-2015"*, Saint Petersburg, Russia (pp. 434–442).